American University of Armenia

Capstone project

Race From Face

Author:

Arpi Hunanyan

Supervisor:

Varduhi Yeghiazaryan

A project submitted in fulfillment of the requirements

for the degree of BS in Data Science in the

Zaven & Sonia Akian College of Science and Engineering

May 19, 2022

Abstract

Applications of race classification are one of the base components of security and defense industries. This project attempts the race classification problem from facial images using transfer learning. All pre-trained models are from Keras Applications. The main pre-trained models are DenseNet121, ResNet50, MobileNetV3Large, and EfficientNetV2B2. These models are pre-trained on the ImageNet data set.

This project uses the FairFace data set which includes several race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. The authors of FairFace collected the data set from the YFCC-100M Flickr data set and then labeled the images with race, gender, and age group. For some of the experiments, we use the MaskTheFace algorithm for adding surgical masks to the images from the FairFace data set.

All convolutional neural network models include pre-trained Keras models and a classification layer. Firstly, the models are trained on the FairFace data set. Afterwards, the MobileNetV3Large is chosen as the most efficient model (in terms of the trade-off between accuracy and memory requirements) based on a comparative experiment. We evaluate that model on the data set including masked face images. Finally, the model is trained on the masked data set to compare its performance in two setups: when it sees only images without masks and when it sees images with and without masks.

Contents

1 Introduction

Physical characteristics like bone structure, skin, hair, and eye color divide people into race groups. Nowadays, the streets of modern cities and towns have cameras that record the view for 24 hours a day. This provides large data sets of image and video recordings.

In parallel, computer vision, computer graphics, and machine learning algorithms have advanced in face recognition and classification tasks. Furthermore, because of these developments, facial analysis has gained significant potential in real-world applications such as security, defense, surveillance, human–computer interaction, biometric identification, and others. Face processing and facial recognition have a wide range of potential applications related to marketing, security, and safety [27].

These data sets and advanced algorithms can be used for solving different complex problems with more accuracy and less complexity. The problems can include face recognition, face verification, gender recognition, age recognition, facial expression recognition, race classification, etc.

The application of race classification can be widely used in security services. Finding someone from a large data set of images and videos is a complex problem. Still, this complexity can be decreased if large data sets are reduced to smaller ones, including only the information of one race type.

There are seven commonly accepted race groups: East Asian, Southeast Asian,

White/Caucasian, Latino/Hispanic, Black, Indian, and Middle Eastern [11]. Even though several works have attempted the problem of race classification in the past, they typically identify one or two race groups. Documented in the early 20th century, the "cross race effect" refers to humans being better at recognizing faces of their own race than other races. Machine learning algorithms work similarly, which makes it critical to include equally balanced data sets for training purposes [8].

Local features such as eyes, lashes, eyebrows, cheeks, nose, and others are comparably good indicators for race [8, 3, 2]. Convolutional neural networks are capable of handling the construction of these local features inside the neural network [11, 8]. However, skin color is not considered to be a good indicator since under different lighting conditions different people can gain similar skin color [11, 8, 2, 5]. There is a certain tendency to solve this problem by using global features (e.g. body movement, speech patterns) from video recordings [11, 31, 28].

This project attempts race classification from face images. The remainder of this work is organized as follows. In Section 2 related work from the scientific literature is presented that touches the problem of race classification. Section 3 describes the data set, and Section 4 describes the structure of the model used to solve the race classification problem. Section 5 discusses experiments and results. Finally, Section 6 concludes this work and proposes future work. Additionally, Appendix A lists the courses and tutorials taken by this author to prepare for this project.

2 Related Work

This section introduces related work on methodologies used for the problem of race classification from face images. The existing literature can be divided into categories. Based on the different classification approaches, we can separate

- machine learning (ML) classical approaches [3, 28, 27, 17, 15, 6, 26, 23, 22]
- and more recently popular deep learning approaches [8, 2, 30, 31].

The overall procedure can include

- only a classification algorithm or
- initial feature extraction, followed by classification.

Feature extraction, in turn, can be local—only using facial features—or global, potentially extending to other body parts or additional cues [2].

The number of classes considered varies depending on the classification approach. For instance, [23] considers Asian vs Non-Asian classes, [29] uses Caucasian, African-American and Asian classes. Many approaches use the k-nearest neighbors algorithm (KNN) from classical ML with different formulas for the distance metric gaining accuracy in the range from 34% to 99% [28, 26]. Another classical machine learning approach is the support vector machine (SVN) with similar accuracy results [3, 2, 27, 17, 22].

It is critical to notice that it has been possible to reach accuracy higher than 98% with classical machine learning approaches with only one class—Asian and Non-Asian [23]—or two classes, such as White and Asian [6, 26]. Furthermore, models with classical machine learning classifiers include additional feature extraction for enhancing the input of the classifier and gaining high accuracy [2].

A deep learning classification algorithm choice is a convolutional neural network (CNN) with better performance on two or three classes [8]. Any deep learning model training from scratch should be trained on a large data set [18] to achieve high accuracy. For better performance, data sets should also be balanced [8]. As it is difficult to find or construct a data set that is both large and balanced, deep learning approaches to date usually rely on three to four classes. However, this still doesn't cover all race groups [16].

Transfer learning, which uses a pre-trained base model for new problems on smaller data sets, has better performance than learning from scratch on image classification problems [18]. Examples of pre-trained models used include VGG16, ResNet, MobileNet, Xception, and others [18]. Recent studies demonstrate that convolutional neural networks can become substantially deeper and more accurate. DenseNet is one of the CNN model examples where a deeper network controls the vanishing gradient problem and substantially reduces the number of parameters which in turn reduces the time of training phase execution [18, 16].

3 Data Set

Face classification data sets can usually have different issues. In machine learning and deep learning, data sets should be large and balanced [11, 8, 3, 30, 18, 16]. Various open-source face data sets, as accumulated in distinct ways, have different problems related to the data distribution.

Firstly, most data sets include 50% White/Caucasian race and much smaller shares of the other races. This scattering creates an unbalance. In the end, the deep learning or machine learning model will classify the White/Caucasian race more accurately than the other races [2, 16].

The second issue related to data sets can be gender inequality. Most data sets include more images of men than women [16]. Also, most images are taken as profile images with professional cameras. Because of that, ML models trained on that kind of data sets in their applications can have unfairness towards non-professional and non-profile images [16].

Another approach is to accumulate data sets from photographs of public figures such as politicians or celebrities. However, this includes a bias as well, since politicians may be in a specific age group and actors and singers may be more attractive than typical faces [16].

Another tendency is to accumulate data sets just by searching for key phrases like “Asian boy” which in turn produces literally Asian boys’ images. But in a

FIGURE 1: Typical image examples of "Asian Boy", celebrities with makeup and filters, and a politician in a specific age group.

real-life scenario, it does not work the same way. So, ML models trained on data sets accumulated in this way will not work without bias and will not produce good results in their applications. Sample images with the above-mentioned problems are illustrated in Figure 1.

In 2021, the authors of the FairFace data set tried to solve the problems mentioned above. They attempted to construct a data set with a balanced distribution of race, gender, and age groups. The data set is collected from the YFCC-100M

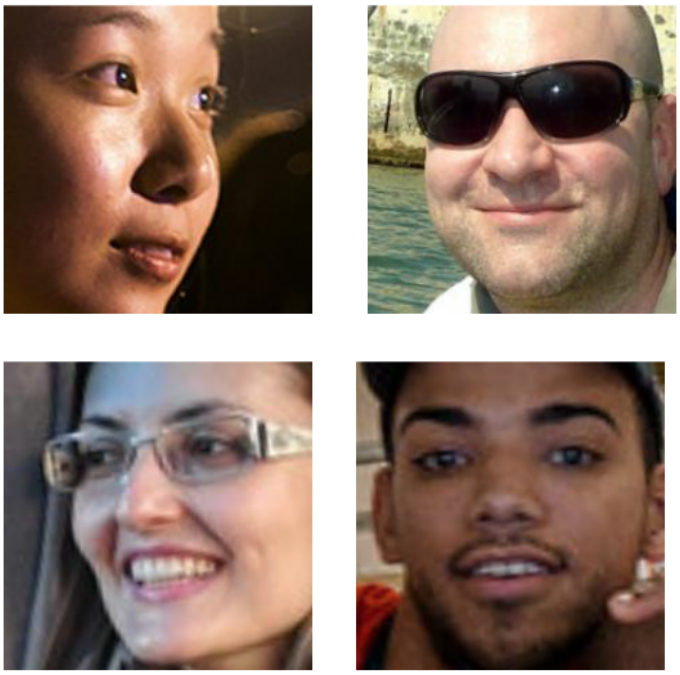

FIGURE 2: Examples from the FairFace data set with profile/non-profile images. Some of them include glasses.

Flickr data set, which includes 100 million media objects, videos, and photos. All the images in the new data set are labeled with race, gender, and age. FairFace includes a training data set with 86,744 images and a validation data set with 10,954 images.

This data set includes seven race groups: White, Black, Indian, East Asian, Latino, Middle Eastern, and Southeast Asian. Furthermore, this data set consists of all face images detected from the dlib's CNN-based detector algorithm [16]. Examples from the FairFace data set are illustrated in Figure 2.

4 Methodology

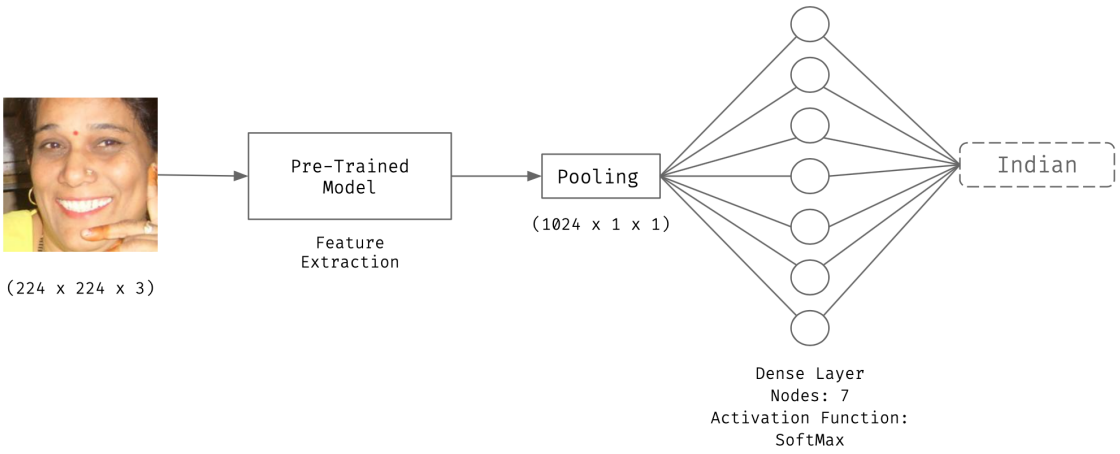

FIGURE 3: The structure of the classification model with the pre-trained model, pooling layer, and dense layer for classification.

Most studies solve the race classification problem from scratch (not using pre-trained models) [8, 2, 16] and all models are trained on data sets that do not include images of faces wearing masks [8, 3, 2, 28, 27, 17, 15, 6, 26, 23, 22, 30, 31].

This project aims to find the best pre-trained model for race classification using a transfer learning approach. Then the goal is to compare models: when they see only images without masks and when they see images with and without masks.

All models used in this project include two main parts—the pre-trained model from Keras Applications [7] and the classification component for classifying race groups. The latter comprises two layers: a pooling layer and a dense layer with seven nodes with the SoftMax activation function for classification. The structure of the model is demonstrated in Figure 3.

All pre-trained models come from Keras Applications [7]. They are trained on ImageNet data set [10] which includes 1.4 million images. Because 17% of that data set is face images, ImageNet can be considered similar to the FairFace data set. So, pre-trained Keras Applications layers can be in training mode during the training phase on the FairFace data set [18, 10].

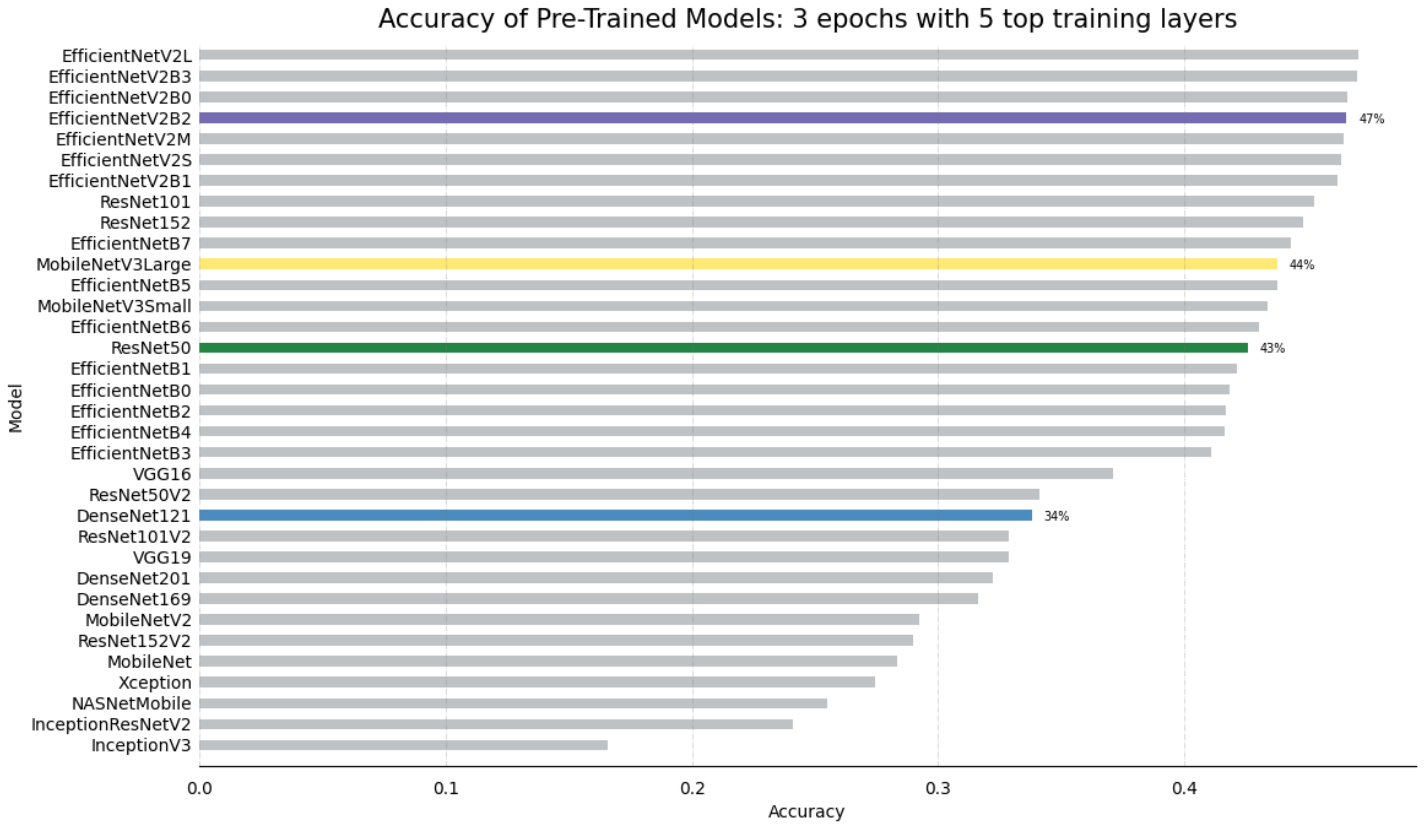

FIGURE 4: Accuracy results of all Keras Applications trained on the FairFace data set. All pre-trained model layers are in inference mode, only the top 5 layers including the classification layer are in training mode. The highlighted lines correspond to the models used in our experiments, i.e. DenseNet121, ResNet50, MobileNetV3Large, and EfficientNetV2B2.

5 Experiments and Results

5.1 Model Choice for Transfer Learning

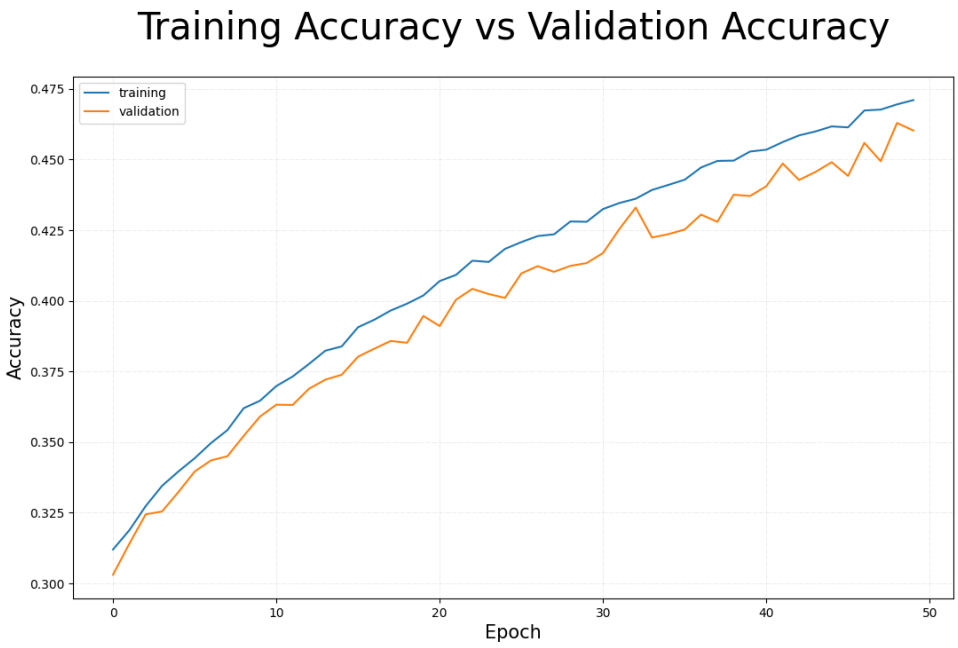

FIGURE 5: Fine-tuning phase. The MobileNetV3Large model as a pre-trained model trained on FairFace with all pre-trained model layers in training mode, when learning rate equals 10^{-7} with 50 epochs.

The first pre-trained model is DenseNet121 [14]. DenseNet solves vanishing gradient problems by reusing tensor outputs of previous layers. This model's pre-trained part is in inference mode, and only the classification layer is in training mode for the first execution. The model with a learning rate of 0.01 in 100 epochs gains accuracy near to 38%, which is very similar to the training accuracy of 33%. Afterwards, all the pre-trained model layers including the classification layer are in

training mode with a learning rate of 10^{-7} in 5 epochs. The model gains accuracy equal to 40% while training accuracy is 41%. These results show that the model can be trained in more epochs since significant over-learning is not detected.

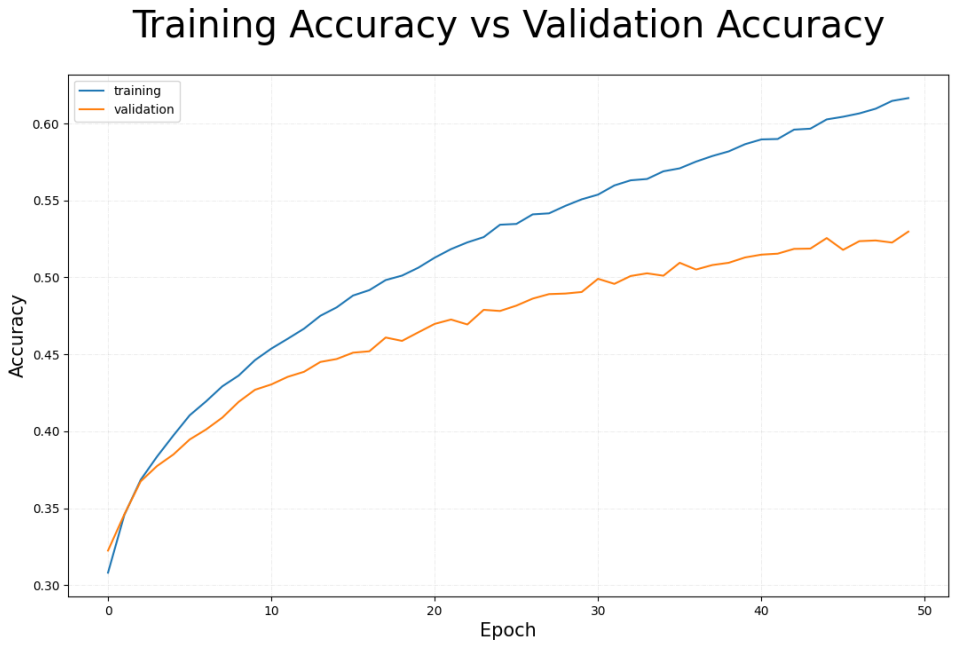

FIGURE 6: Fine-tuning phase. The ResNet50 model as a pre-trained model trained on FairFace with all pre-trained model layers in training mode, when learning rate equals 10^{-7} with 50 epochs.

Keras Applications includes other pre-trained models. Because of the structure, one of them can suit more for this specific image classification task [7]. For this reason, all Keras pre-trained models are trained with a learning rate equal to 0.01 in three epochs. The top five layers are in training mode and the others in inference mode. The results are demonstrated in Figure 4. As it shows, the family of EfficientNet gains top accuracy from 30% to 40%.

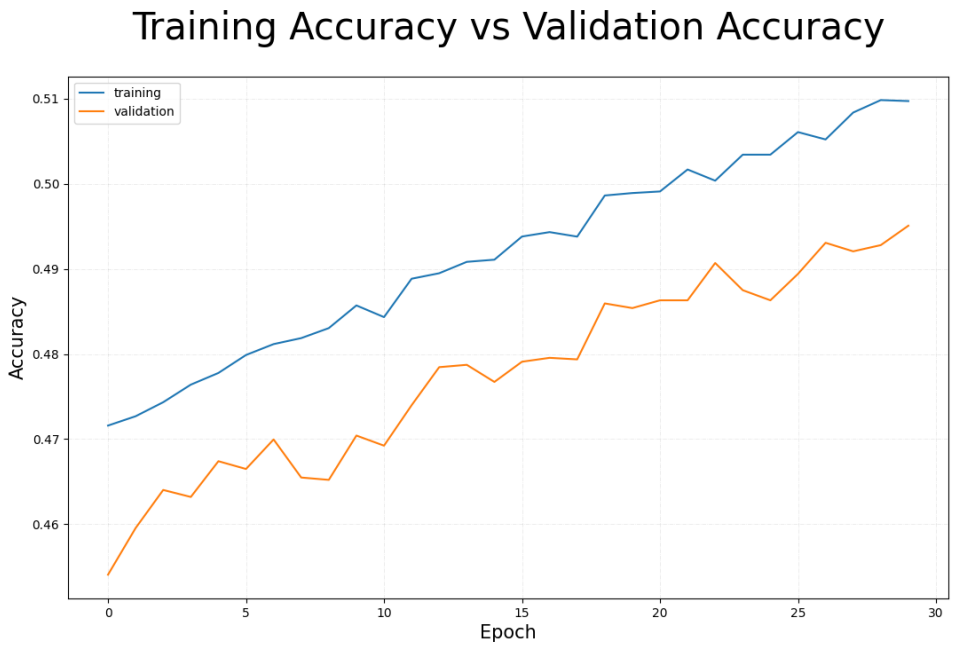

FIGURE 7: The continuation of the fine-tuning phase of MobileNetV3Large model as a pre-trained model when learning rate equals 10^{-7} with another 30 epochs.

EfficientNet uniformly scales all dimensions of depth/width/resolution using a compound coefficient [24]. In particular, EfficientNetV2, introduced in 2021, takes a small input size and uses a small kernel size, making fewer kernel parameters. The structure of the model is very deep (e.g. EfficientNetV2B2 includes 327 layers). Furthermore, the pre-trained EfficientNetV2B2 model generates the weights in progressive learning. In other words, at the beginning of the training phase, they use a small image with weak regularization, but in the end, they use a large image with strong regularization, which prevents over-learning [24].

Since everything comes with a tradeoff, in EfficientNet, data movement between

layers is a very expensive operation [19]. The number of channels increases from one layer to another to boost the overall capacity. Hardware accelerators, like graphics processing units (GPUs), are usually designed to work with models where data movement is a relatively small component of overall performance.

As a result of this problem, the project includes only the training phase of the model, with EfficientNetV2B2 having only the last ten layers in training mode. The model is trained in two hundred epochs with a learning rate of 0.01. It gains accuracy equal to 48%.

The next pre-trained model is chosen from Keras Applications by order of accuracy gained in the selection experiment. The model is MobileNetV3Large which comes after EfficientNet's. The structure of the MobileNetV3 model is based on the input and output layers connected with residual connection only if their numbers of channels correspond to each other [13]. This kind of structure prevents the degradation problem. The degradation problem occurs if the depth of the network leads to a decrease in the performance on both the validation and training data. Also, MobileNetV3 uses depth-wise separable convolution instead of traditional convolution, which decreases the number of parameters and computation time. This makes MobileNetV3 a very efficient model.

The paper titled "Searching for MobileNetV3" reveals two models: MobileNetV3 large and MobileNetV3 small [13]. The first one is for high resource use cases and

the second for low resources. Keras includes both models [7]. The comparison of all Keras Applications models in Figure 4 shows that MobileNetV3Large performs better on the FairFace data set than MobileNetV3Small.

The model containing the pre-trained MobileNetV3Large is trained. All pre-trained model layers are in inference mode, and only the ten last layers are in training mode including the classification layer. The model has a learning rate equal to 0.01 and is trained for 50 epochs. Then, the same model is fine-tuned with a learning rate of 10^{-7} with 50 epochs. The overall result gains 48% accuracy.

With the same hyperparameters, ResNet50 [25] as a pre-trained model is demonstrated afterwards. In the best model selection experiment, these two pre-trained models gained the highest scores after the EfficientNet family.

It appears that pre-trained ResNet50 gains accuracy equal to 52%. However, from the plots in Figures 5 and 6, we can see that ResNet50 tends to overlearn while MobileNetV3Large gains similar accuracy on both validation and training data sets. As can been seen from Figure 5, the training and validation accuracy curves grow in parallel and don't divert from each other as in Figure 6.

MobileNetV3Large pre-trained model with updated weights is fine-tuned with a learning rate of 10^{-7} for another 30 epochs to achieve better performance. Results show that this last update achieves accuracy of nearly 50%. The results are shown in the plot inside Figure 7.

5.2 Experiments on Masked Images

FIGURE 8: FairFace image examples.

The second aim of this project is to compare the accuracy of two models. The first is the existing model with pre-trained MobileNetV3 having nearly 50% accuracy.

The other is the same model after fine-tuning on the same data set of face images but with added masks in 30 epochs with a learning rate of 10^{-7}.

FIGURE 9: FairFace images, after MaskTheFace algorithm [1].

Since it is difficult to find a data set with mask images labeled with race groups, this project uses an algorithm to wear masks on some FairFace images.

The MaskTheFace [1] algorithm is a computer-vision-based script that modifies face images to simulate the wearing of face masks. Firstly, it detects the face landmarks, estimates mask key position and face tilt angle and then selects the right template based on the face tilt. Afterwards, it can add the mask based on the mask key position. Finally, it overlays the mask with adjusted brightness.

In Figure 8 some sample images from the FairFace data set are demonstrated. And in Figure 9 are the same images after the MaskTheFace algorithm.

The model with pre-trained MobileNetV3Large evaluates metrics on masked images and gains 26% accuracy. Afterwards, the same model is fine-tuned on the masked image with a 10^{-7} learning rate for 30 epochs. The results show that it achieves nearly 50% accuracy again. This result indicates that the model keeps its initial performance on the masked data set by training 30 epochs on the data set consisting of face images with masks.

6 Conclusion and Future Work

This project presents an approach based on transfer learning to classify race groups from face images. Pre-trained models are used for feature extraction and a dense layer with SoftMax activation function is used for classification. The models are trained on the FairFace data set, and the one with the best performance is chosen to compare the results on the same data set but with masked faces. The evaluation of the same model once trained on masked images and once without masks indicates that similar results can be accomplished by only fine-tuning the model without masks with 30 epochs on the masked data set.

As the results show, many models with pre-trained Keras Applications do not overlearn during the training and tuning phases. Furthermore, these models can gain high performance by modifying different hyper-parameters. This approach can be used in the future to improve model accuracy.

This project uses an additional dense layer with a SoftMax activation function for the classification part. Still, that layer can be replaced with a classical machine learning classification algorithm, such as KNN or SVN, which will also improve the whole model's performance. The models demonstrated in this project take only one input: a face image, but in the future, they can take face images and some feature images like eyes, nose, eyebrows, and forehead, which also have the potential to improve the performance of the model.

A Courses Taken in Preparation of the Project

To accomplish this project, I needed to cover Coursera's "Deep Learning course" [20] for theoretical deep learning knowledge. For the practical component of deep learning, I took "Introduction to Deep Learning in Python" [4], "Advanced Deep Learning with Keras" [9] in DataCamp. Also, I read Keras tutorials for implementing Keras Applications [7]. Additionally, I needed to work with images, I took "Image Processing with Keras in Python" [21] and "Image Processing in Python" [12] DataCamp courses.

References

[1] Aqeel Anwar and Arijit Raychowdhury. Masked face recognition for secure authentication, 2020.

[2] Inzamam Anwar and Naeem Ul Islam. Learned features are better for ethnicity classification. *arXiv preprint arXiv:1709.07429*, 2017.

[3] Fabiola Becerra-Riera, Nelson Méndez Llanes, Annette Morales-González, Heydi Méndez-Vázquez, and Massimo Tistarelli. On combining face local appearance and geometrical features for race classification. In *Iberoamerican Congress on Pattern Recognition*, pages 567–574. Springer, 2018.

[4] Dan Becker. Introduction to deep learning in python, 2022.

[5] Isabelle Bülthoff, Wonmo Jung, Regine GM Armann, and Christian Wallraven. Predominance of eyes and surface information for face race categorization. *Scientific reports*, 11(1):1–9, 2021.

[6] Hengxin Chen, Mingqi Gao, Karl Ricanek, Weiliang Xu, and Bin Fang. A novel race classification method based on periocular features fusion. *International Journal of Pattern Recognition and Artificial Intelligence*, 31(08):1750026, 2017.

[7] Francois Chollet et al. Keras, 2015.

[8] AS Darabant, D Borza, and R Danescu. Recognizing human races through machine learning—a multi-network, multi-features study. mathematics 2021, 9, 195, 2021.

[9] Zachary Deane-Mayer. Advanced deep learning with keras, 2019.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[11] Siyao Fu, Haibo He, and Zeng-Guang Hou. Learning race from face: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2483–2509, 2014.

[12] Rebeca Gonzalez. Image processing in python, 2021.

[13] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[15] Sasan Karamizadeh and Shahidan M Abdullah. Race classification using gaussian-based weight k-nn algorithm for face recognition. *Journal of Engineering Research*, 6(2):103–121, 2018.

[16] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[17] Hui Lin, Huchuan Lu, and Lihe Zhang. A new automatic recognition system of gender, age and ethnicity. In *2006 6th world congress on intelligent control and automation*, volume 2, pages 9988–9991. IEEE, 2006.

[18] Pedro Marcelino. Transfer learning from pre-trained models. *Towards Data Science*, 10:23, 2018.

[19] neuralmagic. The challenges of efficientnets and the way forward. https://neuralmagic.com/blog/the-challenges-of-efficientnets-and-the-way-forward/, May 2022.

[20] Andrew Ng. Coursera deep learning, 2020.

[21] Ariel Rokem. Image processing with keras in python, 2020.

[22] S Md Mansoor Roomi, SL Virasundarii, S Selvamegala, S Jeevanandham, and D Hariharasudhan. Race classification based on facial features. In *2011 third*

national conference on computer vision, pattern recognition, image processing and graphics, pages 54–57. IEEE, 2011.

[23] Mezzoudj Saliha, Behloul Ali, and Seghir Rachid. Towards large-scale face-based race classification on spark framework. *Multimedia Tools and Applications*, 78(18):26729–26746, 2019.

[24] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021.

[25] Sasha Targ, Diogo Almeida, and Kevin Lyman. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*, 2016.

[26] George Toderici, Sean M O'malley, George Passalis, Theoharis Theoharis, and Ioannis A Kakadiaris. Ethnicity-and gender-based subject retrieval using 3-d face-recognition techniques. *International Journal of Computer Vision*, 89(2):382–391, 2010.

[27] Eone Etoua Oscar Vianney, Tapamo Kenfack Hippolyte Michel, Mboule Ebele Brice Auguste, Mbietieu Amos Mbietieu, and Essuthi Essoh Serge Leonel. Race recognition using enhanced local binary pattern. In *Pan-African Artificial Intelligence and Smart Systems Conference*, pages 120–136. Springer, 2021.

[28] Shakhawan Hares Wady and Hawkar Omar Ahmed. Ethnicity identification based on fusion strategy of local and global features extraction. *International Journal of Multidisciplinary and Current Research*, 4(2):200–205, 2016.

[29] Yiting Xie, Khoa Luu, and Marios Savvides. A robust approach to facial ethnicity classification on large scale face databases. In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 143–149. IEEE, 2012.

[30] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020.

[31] De Zhang, Yunhong Wang, and Bir Bhanu. Ethnicity classification based on gait using multi-view fusion. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 108–115. IEEE, 2010.